DOCUMENT RESUME

ED 406 441 TM 026 420

AUTHOR Matlock-Hetzel, Susan
TITLE Basic Concepts in Item and Test Analysis.
PUB DATE 23 Jan 97
NOTE 22p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 23-25, 1997).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Difficulty Level; *Distractors (Tests); Elementary Secondary Education; *Item Analysis; *Multiple Choice Tests; *Norm Referenced Tests; Test Construction; Test Items
IDENTIFIERS Item Discrimination (Tests); *Test Analysis

ABSTRACT

When norm-referenced tests are developed for instructional purposes, to assess the effects of educational programs, or for educational research purposes, it can be very important to conduct item and test analyses. These analyses can evaluate the quality of items and of the test as a whole. Such analyses can also be employed to revise and improve both items and the test as a whole. However, some best practices in item and test analysis are too infrequently used in actual practice. This paper summarizes recommendations for item and test analysis practices as are reported in commonly used textbooks. These practices are determination of item difficulty, item discrimination, and item distractors. Item difficulty is simply the percentage of students taking the test who answered the item correctly. The larger the percentage getting the item right, the easier the item. A good test item discriminates between those who do well on the test and those who do poorly. The item discrimination index and discrimination coefficients can be computed to determine the discriminating power of an item. In addition, analyzing the distractors (incorrect alternatives) is useful in determining the relative usefulness of the decoy items, which should be modified if students consistently fail to select certain multiple choice alternatives. These techniques can help provide empirical information about how tests are performing in real test situations. (Contains 7 tables and 13 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 406 441

Running Head: BASIC CONCEPTS

Basic Concepts in Item and Test Analysis

Susan Matlock-Hetzel

Texas A&M University 77843-4225

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY

SUSAN MATLOCK-HETZEL

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

TM026420

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, Tx, January 23, 1997.

Abstract

When norm-referenced tests are developed for instructional purposes, to assess the effects of educational programs, or for educational research purposes, it can be very important to conduct item and test analyses. These analyses evaluate the quality of the items and of the test as a whole. Such analyses can also be employed to revise and improve both items and the test as a whole. However, some best practices in item and test analysis are too infrequently used in actual practice. The purpose of the present paper is to summarize the recommendations for item and test analysis practices, as these are reported in commonly-used measurement textbooks (Crocker & Algina, 1986; Gronlund & Linn, 1990; Pedhazur & Schemlkin, 1991; Sax, 1989; Thorndike, Cunningham, Thorndike, & Hagen, 1991).

Basic Concepts in Item and Test Analysis

Making fair and systematic evaluations of others' performance can be a challenging task. Judgments cannot be made solely on the basis of intuition, haphazard guessing, or custom (Sax, 1989). Teachers, employers, and others in evaluative positions use a variety of tools to assist them in their evaluations. Tests are tools that are frequently used to facilitate the evaluation process. When norm-referenced tests are developed for instructional purposes, to assess the effects of educational programs, or for educational research purposes, it can be very important to conduct item and test analyses.

Test analysis examines how the test items perform as a set. Item analysis "investigates the performance of items considered individually either in relation to some external criterion or in relation to the remaining items on the test" (Thompson & Levitov, 1985, p. 163). These analyses evaluate the quality of items and of the test as a whole. Such analyses can also be employed to revise and improve both items and the test as a whole.

However, some best practices in item and test analysis are too infrequently used in actual practice. The purpose of the present paper is to summarize the recommendations for item and test analysis practices, as these are reported in commonly-used measurement textbooks (Crocker & Algina, 1986; Gronlund & Linn, 1990; Pedhazur & Schemlkin, 1991; Sax, 1989; Thorndike,

Cunningham, Thorndike, & Hagen, 1991). These tools include item difficulty, item discrimination, and item distractors.

Item Difficulty

Item difficulty is simply the percentage of students taking the test who answered the item correctly. The larger the percentage getting an item right, the easier the item. The higher the difficulty index, the easier the item is understood to be (Wood, 1960). To compute the item difficulty, divide the number of people answering the item correctly by the total number of people answering item. The proportion for the item is usually denoted as p and is called item difficulty (Crocker & Algina, 1986). An item answered correctly by 85% of the examinees would have an item difficulty, or p value, of .85, whereas an item answered correctly by 50% of the examinees would have a lower item difficulty, or p value, of .50.

A p value is basically a behavioral measure. Rather than defining difficulty in terms of some intrinsic characteristic of the item, difficulty is defined in terms of the relative frequency with which those taking the test choose the correct response (Thorndike et al, 1991). For instance, in the example below, which item is more difficult?

1. Who was Boliver Scagnasty?
2. Who was Martin Luther King?

One cannot determine which item is more difficult simply by

reading the questions. One can recognize the name in the second question more readily than that in the first. But saying that the first question is more difficult than the second, simply because the name in the second question is easily recognized, would be to compute the difficulty of the item using an intrinsic characteristic. This method determines the difficulty of the item in a much more subjective manner than that of a p value.

Another implication of a p value is that the difficulty is a characteristic of both the item and the sample taking the test. For example, an English test item that is very difficult for an elementary student will be very easy for a high school student. A p value also provides a common measure of the difficulty of test items that measure completely different domains. It is very difficult to determine whether answering a history question involves knowledge that is more obscure, complex, or specialized than that needed to answer a math problem. When p values are used to define difficulty, it is very simple to determine whether an item on a history test is more difficult than a specific item on a math test taken by the same group of students.

To make this more concrete, take into consideration the following examples. When the correct answer is not chosen (p = 0), there are no individual differences in the "score" on that item. As shown in Table 1, the correct answer C was not chosen by either the upper group or the lower group. (The upper group

and lower group will be explained later.) The same is true when everyone taking the test chooses the correct response as is seen in Table 2. An item with a p value of .0 or a p value of 1.0 does not contribute to measuring individual differences, and this is almost certain to be useless. Item difficulty has a profound effect on both the variability of test scores and the precision with which test scores discriminate among different groups of examinees (Thorndike et al, 1991). When all of the test items are extremely difficult, the great majority of the test scores will be very low. When all items are extremely easy, most test scores will be extremely high. In either case, test scores will show very little variability. Thus, extreme p values directly restrict the variability of test scores.

In discussing the procedure for determining the minimum and maximum score on a test, Thompson and Levitov (1985) stated that

> items tend to improve test reliability when the percentage of students who correctly answer the item is halfway between the percentage expected to correctly answer if pure guessing governed responses and the percentage (100%) who would correctly answer if everyone knew the answer. (pp. 164-165)

For example, many teachers may think that the minimum score on a test consisting of 100 items with four alternatives each is 0, when in actuality the theoretical floor on such a test is 25. This is the score that would be most likely if a student answered

every item by guessing (e.g., without even being given the test booklet containing the items).

Similarly, the ideal percentage of correct answers on a four-choice multiple-choice test is not 70-90%. According to Thompson and Levitov (1985), the ideal difficulty for such an item would be halfway between the percentage of pure guess (25%) and 100%, (25% + {(100% - 25%)/2}. Therefore, for a test with 100 items with four alternatives each, the ideal mean percentage of correct items, for the purpose of maximizing score reliability, is roughly 63%. Tables 3, 4, and 5 show examples of items with p values of roughly 63%.

Item Discrimination

If the test and a single item measure the same thing, one would expect people who do well on the test to answer that item correctly, and those who do poorly to answer the item incorrectly. A good item discriminates between those who do well on the test and those who do poorly. Two indices can be computed to determine the discriminating power of an item, the item discrimination index, D, and discrimination coefficients.

Item Discrimination Index, D

The method of extreme groups can be applied to compute a very simple measure of the discriminating power of a test item. If a test is given to a large group of people, the discriminating power of an item can be measured by comparing the number of

people with high test scores who answered that item correctly with the number of people with low scores who answered the same item correctly. If a particular item is doing a good job of discriminating between those who score high and those who score low, more people in the top-scoring group will have answered the item correctly.

In computing the discrimination index, D, first score each student's test and rank order the test scores. Next, the 27% of the students at the top and the 27% at the bottom are separated for the analysis. Wiersma and Jurs (1990) stated that "27% is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analysis" (p. 145). There need to be as many students as possible in each group to promote stability, at the same time it is desirable to have the two groups be as different as possible to make the discriminations clearer. According to Kelly (as cited in Popham, 1981) the use of 27% maximizes these two characteristics. Nunnally (1972) suggested using 25%.

The discrimination index, D, is the number of people in the upper group who answered the item correctly minus the number of people in the lower group who answered the item correctly, divided by the number of people in the largest of the two groups. Wood (1960) stated that

when more students in the lower group than in the upper

> group select the right answer to an item, the item actually has negative validity. Assuming that the criterion itself has validity, the item is not only useless but is actually serving to decrease the validity of the test. (p. 87)

The higher the discrimination index, the better the item because such a value indicates that the item discriminates in favor of the upper group, which should get more items correct, as shown in Table 6. An item that everyone gets correct or that everyone gets incorrect, as shown in Tables 1 and 2, will have a discrimination index equal to zero. Table 7 illustrates that if more students in the lower group get an item correct than in the upper group, the item will have a negative D value and is probably flawed.

A negative discrimination index is most likely to occur with an item covers complex material written in such a way that it is possible to select the correct response without any real understanding of what is being assessed. A poor student may make a guess, select that response, and come up with the correct answer. Good students may be suspicious of a question that looks too easy, may take the harder path to solving the problem, read too much into the question, and may end up being less successful than those who guess. As a rule of thumb, in terms of discrimination index, .40 and greater are very good items, .30 to .39 are reasonably good but possibly subject to improvement, .20

to .29 are marginal items and need some revision, below .19 are considered poor items and need major revision or should be eliminated (Ebel & Frisbie, 1986).

Discrimination Coefficients

Two indicators of the item's discrimination effectiveness are point biserial correlation and biserial correlation coefficient. The choice of correlation depends upon what kind of question we want to answer. The advantage of using discrimination coefficients over the discrimination index (D) is that every person taking the test is used to compute the discrimination coefficients and only 54% (27% upper + 27% lower) are used to compute the discrimination index, D.

Point biserial. The point biserial (r_{pbis}) correlation is used to find out if the right people are getting the items right, and how much predictive power the item has and how it would contribute to predictions. Henrysson (1971) suggests that the r_{pbis} tells more about the predictive validity of the total test than does the biserial r, in that it tends to favor items of average difficulty. It is further suggested that the r_{pbis} is a combined measure of item-criterion relationship and of difficulty level.

Biserial correlation. Biserial correlation coefficients (r_{bis}) are computed to determine whether the attribute or attributes measured by the criterion are also measured by the

item and the extent to which the item measures them. The r_{bis} gives an estimate of the well-known Pearson product-moment correlation between the criterion score and the hypothesized item continuum when the item is dichotomized into right and wrong (Henrysson, 1971). Ebel and Frisbie (1986) state that the r_{bis} simply describes the relationship between scores on a test item (e.g., "0" or "1") and scores (e.g., "0", "1",..."50") on the total test for all examinees.

Distractors

Analyzing the distractors (e.i., incorrect alternatives) is useful in determining the relative usefulness of the decoys in each item. Items should be modified if students consistently fail to select certain multiple choice alternatives. The alternatives are probably totally implausible and therefore of little use as decoys in multiple choice items. A discrimination index or discrimination coefficient should be obtained for each option in order to determine each distractor's usefulness (Millman & Greene, 1993). Whereas the discrimination value of the correct answer should be positive, the discrimination values for the distractors should be lower and, preferably, negative. Distractors should be carefully examined when items show large positive D values. When one or more of the distractors looks extremely plausible to the informed reader and when recognition of the correct response depends on some extremely subtle point,

BEST COPY AVAILABLE

it is possible that examinees will be penalized for partial knowledge.

Thompson and Levitov (1985) suggested computing reliability estimates for a test scores to determine an item's usefulness to the test as a whole. The authors stated, "The total test reliability is reported first and then each item is removed from the test and the reliability for the test less that item is calculated" (Thompson & Levitov, 1985, p.167). From this the test developer deletes the indicated items so that the test scores have the greatest possible reliability.

Summary

Developing the perfect test is the unattainable goal for anyone in an evaluative position. Even when guidelines for constructing fair and systematic tests are followed, a plethora of factors may enter into a student's perception of the test items. Looking at an item's difficulty and discrimination will assist the test developer in determining what is wrong with individual items. Item and test analysis provide empirical data about how individual items and whole tests are performing in real test situations.

References

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Ebel, R.L., & Frisbie, D.A. (1986). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N.E., & Linn, R.L. (1990). Measurement and evaluation in teaching (6th ed.). New York: MacMillan.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R.L. Thorndike (Ed.), Educational Measurement (p. 141). Washington DC: American Council on Education.

Millman, J., & Greene, J. (1993). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), Educational measurement (pp. 335-366). Phoenix, AZ: Oryx Press.

Nunnally, J.C. (1972). Educational measurement and evaluation (2nd ed.). New York: McGraw-Hill.

Pedhazur, E.J., & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

Popham, W.J. (1981). Modern educational measurement. Englewood Cliff, NJ: Prentice-Hall.

Sax, G. (1989). Principles of educational and psychological measurement and evaluation (3rd ed.). Belmont, CA: Wadsworth.

Thompson, B., & Levitov, J.E. (1985). Using microcomputers

to score and evaluate test items. Collegiate Microcomputer, 3, 163-168.

Thorndike, R.M., Cunningham, G.K., Thorndike, R.L., & Hagen, E.P. (1991). Measurement and evaluation in psychology and education (5th ed.). New York: MacMillan.

Wiersma, W. & Jurs, S.G. (1990). Educational measurement and testing (2nd ed.). Boston, MA: Allyn and Bacon.

Wood, D.A. (1960). Test construction: Development and interpretation of achievement tests. Columbus, OH: Charles E. Merrill Books, Inc.

Table 1

Minimum Item Difficulty Example Illustrating No Individual Differences

Group	Item Response			
			*	
	A	B	C	D
Upper group	4	5	0	6
Lower group	2	6	0	7

Note. * denotes correct response

Item difficulty: (0 + 0)/30 = .00p

Discrimination Index: (0 - 0)/15 = .00

Table 2

Maximum Item Difficulty Example Illustrating No Individual Differences

Group	Item Response			
			*	
	A	B	C	D
Upper group	0	0	15	0
Lower group	0	0	15	0

Note. * denotes correct response

Item difficulty: (15 + 15)/30 = 1.00p

Discrimination Index: (15-15)/15 = .00

Table 3

Maximum Item Difficulty Example Illustrating Individual Differences

Group	Item Response			
			*	
	A	B	C	D
Upper group	1	0	13	3
Lower group	2	5	5	6

Note. * denotes correct response

Item difficulty: (13 + 5)/30 = .60p

Discrimination Index: (13-5)/15 = .53

Table 4

Maximum Item Difficulty Example Illustrating Individual Differences

Group	Item Response			
			*	
	A	B	C	D
Upper group	1	0	11	3
Lower group	2	0	7	6

Note. * denotes correct response

Item difficulty: (11 + 7)/30 = .60p

Discrimination Index: (11-7)/15 = .267

Table 5

Maximum Item Difficulty Example Illustrating Individual Differences

Group	Item Response			
			*	
	A	B	C	D
Upper group	1	0	7	3
Lower group	2	0	11	6

Note. * denotes correct response

Item difficulty: (11 + 7)/30 = .60p

Discrimination Index: (7 - 11)/15 = .267

Table 6

Positive Item Discrimination Index D

Group	Item Response			
			*	
	A	B	C	D
Upper group	3	2	15	0
Lower group	12	3	3	2

Note. * denotes correct response

74 students took the test

27% = 20(N)

Item difficulty: (15 + 3)/40 = .45p

Discrimination Index: (15 - 3)/20 = .60

Table 7

Negative Item Discrimination Index D

Group	Item Response			
			*	
	A	B	C	D
Upper group	0	0	0	0
Lower group	0	0	15	0

Note. * denotes correct response

Item difficulty: (0 + 15)/30 = .50p

Discrimination Index: (0 - 15)/15 = -1.0